Feeling our way to an analysis of English possessed idioms

Ho Jia Qian♠ Francis Bond,♠ and Dan Flickinger♡
♠ Nanyang Technological University
♡ CSLI, Stanford University

HPSG 2015

1 Introduction

Idiomatic constructions are common in language, both at a type and token level. Despite considerable effort in categorizing and analyzing them (Nunberg et al., 1994; Moon, 1998; Sag et al., 2002, and many, many others) their coverage is still far from complete in lexicons such as WordNet (Fellbaum, 1998a) and grammars such as the English Resource Grammar (Flickinger, 2011).

This paper focuses on possessive idiomatic constructions, prototypically those in which a constituent is modified by a possessive pronoun co-indexed with another constituent (usually the subject). An example is ***wrack one's brains*** "think hard", where the possessor of the brains must be the subject: *I wrack my brains*; *They wrack their brains*. These are interesting theoretically because of the interaction between syntax and semantics. Many languages, even with similar idioms, do not include this possessive expression. For example, the equivalent phrase in Japanese is ***chie-wo shiboru*** "think hard: lit., squeeze knowledge", which contains a verb phrase with a fixed object, but no possessive.

The initial motivation for this research was for machine translation: when translating out of English into Japanese, typically the idiomatic possessive pronoun should be omitted (Bond, 2005). Going the other way, the possessive pronoun must be generated and also agree with the subject. Shallow statistical MT systems often get this wrong. A complete list of these idioms may also be useful for computer-assisted language learning. For example, an English learner can use the materials developed from corpora to understand figurative language, which is a more difficult aspect of language to learn, and to understand how pronouns operate in both literal and figurative English.

2 Analysis

We collected idioms from a variety of sources, including WordNet (Fellbaum, 1998a) and online lexicons like Dictionary.com (2012). We described 514 idioms, which were then broadly classified into ***co-indexed possessive*** and ***separate possessive*** idioms. The major division lies in the possessor being co-indexed with the subject (co-indexed possessive), and those where it is not (separate possessive). There was a further categorization according to their syntactic templates. Here we show only the possessive idioms in Table 1.

For each idiom we then created a rich ***idiom entry***, as in (1). We wrote a definition, listed some examples, and added a paraphrase. For each predicate in the idiom and paraphrase, we determined the sense using WordNet as our sense inventory. For decomposable idioms, we also determined the sense of the metaphorical extension for each component word (marked with *). The decomposability of an idiom was listed using the feature *@type*.

As mentioned, all the idioms were given paraphrases. These were restricted, linked to WordNet and marked with @ in the idiom entries: ***wrack one's brains*** → ***think hard***. Due to the variance in the possessive pronoun, it is hard to extract these paraphrases automatically even using sophisticated methods (Zhang et al., 2006).

Table 1
Types of Co-indexed Possessive Idioms

Structure	Example	Frequency
X_{NP} V_1 X's N_1	lose one's mind	137
X_{NP} V_1 [P_1 X's N_1]	fly off one's handle	40
X_{NP} V_1 X's N_1 [P_1 Y_{NP}]	cast one's lot [with someone/thing]	39
X_{NP} V_1 X's N_1 [P_1 D_1 N_2]	have one's head [in the clouds]	27
X_{NP} V_1 X's N_1 P_1	cry one's eyes out	22
X_{NP} V_1 X's own N_1	blow one's own horn	18
X_{NP} V_1+P_1 X's N_1	pull up one's socks	17
X_{NP} be [P_1 X's N_1]	off one's rocker	13
X_{NP} V_1 X's N_1 [P_1 X's N_2]	scratch one's ear [with one's elbow]	13
X_{NP} V_1 D_1 N_1 [P_1 X's N_2]	a dose [of one's medicine]	10
X_{NP} V_1 X's N_1 A_1	get one's hands dirty	10
X_{NP} V_1 Y_{NP} [P_1 X's N_1]	wind someone [around one's finger]	10
X_{NP} V_1 X's N_1(est)	do one's best	8
X_{NP} V_1 [P_1 X's N_1 [P_2 Y_{NP}]]	pour out one's heart [to someone]	7
X_{NP} aux+neg V_1 X's N_1	not mince one's words	5
X_{NP} V_1 Y_{NP} D_1 N_1 [P_1 X's N_2]	give someone a piece [of one's mind]	4
X_{NP} V_1 R_1 A_1 [P_1 X's N_1]	too big [for one's boots]	3
X_{NP} V_1 [P_1 D_1 N_1 P_2 X's N_2]	by the skin of one's teeth	2
X_{NP} V_1 N_1 [P_1 X's N_2]	have egg [on one's face]	2
X_{NP} V_1 X's N_1 [P_1 X]	have one's wits [about one]	2
X_{NP} V_1 X's N_1 and V_2 N_2	have one's cake and eat it	2
Remainder	let grass grow under one's feet	30
Total		**421**

(1)

Idiom entry	
Index form	rack one's brain
Template	X_{NP} V_1 X's N_1
Definition	to struggle to remember or think of something
V_1	S: (v) rack (torture on the rack)
N_1	S: (n) mind, head, brain, psyche, nous (that which is responsible for one's thoughts, feelings, and conscious brain functions; the seat of the faculty of reason)
$^*V_1$	S: (v) strive, reach, strain (to exert much effort or energy)
$^*N_1$	= N_1
@type	decomposable
@paraphrase	X thinks hard
@V	S: (v) think, cogitate, cerebrate (use or exercise the mind or one's power of reason in order to make inferences, decisions, or arrive at a solution or judgments)
@N	S: (adv) hard (with effort or force or vigor)
comment	***wrack one's brain***

The most common idiom was the co-indexed basic verb phrase idiom (X_{NP} V_1 X's N_1), which constituted 137 out of the 514 idioms. These idioms are typically paraphrased as an intransitive verb plus modifier. However, like many decomposable idioms, the idiomatic noun phrase can be modified, typically with the effect of strengthening or diminishing the idiom: *He wracked his feeble brains*.

After identifying and classifying the idioms, we then analyzed the idioms using the basic approach of Copestake (1994); Sag et al. (2002). The relationship between the words in the idiom is captured using Minimal Recursion Semantics (MRS: Copestake et al., 2005). Special lexical items introduce idiomatic predicates, marked as such in the lexicon. Idioms are treated as bags of predications (Example 2b), with relations between them partially specified and the co-indexation marked by an identity relation *id_rel*. If the semantics of a sentence matches such a specification, then it has the idiomatic reading. This lexically anchored approach allows for considerable syntactic flexibility, and for precise constraints on that flexibility. It takes advantage of the monostratal nature of HPSG. The idioms do not behave unexpectedly with their syntax, but only their semantics — we can therefore place the constraints in the appropriate place: the semantic representation. The analysis was implemented in a computational HPSG of English (the ERG) for the most frequent types.

Miyazaki et al. (1993) suggested that for some idioms we should allow nodes in a semantic hierarchy (so any noun with compatible semantics is allowed). As illustrated in (1), we have linked the predicates in the idiom to their literal meanings and the predicates in their paraphrases to the intended meaning using WordNet synsets.

The idiomatic ***wrack one's brains*** has three elements in the grammar: a lexical entry that introduces ***_brains_n_i***; a lexical entry that introduces ***_wrack_v_i*** and *id_rel* and links the semantic indices appropriately; and an idiom rule that makes sure all the relevant elements are there: the above three predicates and the possessive relation *poss_rel*. To link the subject to the possessor of the object, the identity relation *id_rel* is linked to the external argument (XARG) of the verb (the subject) and to the external argument of the first element of the COMPS list (the determiner of the object) if and only if the two arguments of *id_rel* agree with each other. Minor variations can easily be captured. For example, there are two alternative spellings of ***wrack***: *wrack* and *rack*. If we treat them as having the same meaning, we can have two lexical items with different orthography, but having the same predicate and thus giving rise to the same semantic idiosyncrasy. Another cause of variation is in number: both *I rack my brain* and *I wrack my brains* are attested. In this case, we can underspecify number in the construction and allow both.

(2) a. I_i rack my_i brains. [X_{NP} V_1 X's N_1]

b.
$$\left[\begin{array}{l}\textit{idiom constraint}\\ \text{RELS}\ \left\langle \left[\begin{array}{ll}\textbf{\textit{\_rack\_v\_i\_rel}} & \\ \text{LBL} & \boxed{hv}\\ \text{ARG0} & \boxed{v}\\ \text{ARG1} & \boxed{x}\\ \text{ARG2} & \boxed{z}\end{array}\right], \left[\begin{array}{ll}\textit{id\_rel} & \\ \text{LBL} & \boxed{hv}\\ \text{ARG0} & \boxed{id}\\ \text{ARG1} & \boxed{x}\\ \text{ARG2} & \boxed{y}\end{array}\right], \left[\begin{array}{ll}\textit{poss\_rel} & \\ \text{LBL} & \boxed{hz}\\ \text{ARG0} & \boxed{ps}\\ \text{ARG1} & \boxed{z}\\ \text{ARG2} & \boxed{y}\end{array}\right], \left[\begin{array}{ll}\textbf{\textit{\_brain\_n\_i\_rel}} & \\ \text{LBL} & \boxed{hz}\\ \text{ARG0} & \boxed{z}\end{array}\right] \right\rangle\end{array}\right]$$

3 Results

We implemented the most common types of idioms in the English Resource Grammar. We then used this implementation to look at the amount of lexical, syntactic and semantic variation that was possible.

3.1 Corpus Findings

For lexical variation, we generated paraphrases for all the open predicates, using wordnet. Variants were produced for each predicate by substituting synonyms or direct hyponyms. So ***wrack one's brains*** was expanded to ***wrack one's brain***, ***wrack one's encephalon***.

The original idioms and their variants were then run through the British National Corpus (BNC: Burnard, 2000) in two preliminary studies to determine the kind and frequency of idioms in the current corpus, so as to better understand their semantic and syntactic flexibility.

The first study involved checking for the recall on two idioms *bite one's tongue* and *(w)rack one's brains*. We found all sentences with with either *bite* and *tongue* or *rack* and *brain* and manually selected the 82 sentences containing the idioms. The ERG was able to parse 80% of the sentences, and correctly identified the idiom in 82% of those it could parse. We will try a second run with the full robustness machinery on to parse the remaining 20% . To improve the identification of the idioms in sentences we can parse, we need to improve our treatment of agreement between elements linked with *id_rel*): for example in (3), we do currently consider *their* to agree with *someone*.

(3) *It's all very well telling someone$_i$ to bite their$_i$ tongue and not fight back.*

The second study involved parsing and checking about 100,000 sentences, of which 319 sentences (0.03%) contained idioms and variants. A manual check showed that 76.7% were correctly parsed as idioms. The relatively high percentage validates our methodology, in terms of the rich idiom entries. The ERG implementation allows us to automatically identify these complicated idioms and links to wordnet senses their variants. In our final study we will parse the entire BNC (which we estimate will take around 1,000 parser days.

The most common idiom we found in our random sample of the BNC (almost 20% of the examples) was *shake one's head*, which was often used both literally and figuratively. The relatively high percentage is partly because a single literary text had many examples. This demonstrates how genres can affect the kind of idiom; in this case, *shake one's head* is very common in stories. Since the BNC contains information about genres, future work could examine the relationship between genres and idiom frequency. This would inform English-language learners of what idioms they should learn, depending on their area of interest.

Moreover, since current dictionaries do not list idiom frequency, such a corpus-based study is important not just for enhancing dictionaries, but also for improving translation systems by informing NLP programmers what idioms to focus on.

4 Ongoing work and future work

We are currently still running the idioms and variants over the BNC to identify actual examples of theses idioms. Up until now it has been hard to find these, due to the complicated structure. With idioms implemented in a flexible grammar, they can be identified automatically.

In future work, we will rethink how to mark the idioms in the output semantic representation. Currently, the individual elements are marked as idiomatic. During processing we know which idiom was licensed (as we know which idiom rule applies), but this information is not part of the final MRS. Further, the possessive pronouns are not marked in any way, even though intuitively they are less meaningful than real referential pronouns. Both these issues are also relevant to the separate possession idioms.

Finally, there is a long tail of infrequent idiom types, which still have to be implemented. Some even cross clause boundaries as Richter & Sailer (2009) point out: for example: ***look as though butter wouldn't melt in one's mouth*** "appear innocent".

5 Conclusions

We have developed an analysis of co-indexed possessive idioms in HPSG, and implemented this analysis in a computational grammar, the ERG. We have further linked the relevant lexical semantic predicates to corresponding synsets in WordNet, to accommodate the wider range of lexical variation found in many of these idioms. Using this augmented grammar, we parsed the British National Corpus in order to identify occurrences of these idioms in running text, and manually evaluated the results for a sizeable sample. We are currently working on expanding coverage to more of the long tail of less frequently occurring types, and to further analysis and tuning based on first results from parsing the BNC. As well as the additions to the open-source ERG, we will make the full idiom lexicon, including definitions, examples and links to WordNets freely available under an open license (CC-BY).[1]

References

Bond, Francis. 2005. *Translating the untranslatable: A solution to the problem of generating English determiners* CSLI Studies in Computational Linguistics. CSLI Publications.

Burnard, Lou. 2000. *The British National Corpus users reference guide*. Oxford University Computing Services.

Copestake, Ann. 1994. Representing idioms. Presentation at the HPSG Conference, Copenhagen.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation* 3(4):281-332.

Dictionary.com. 2012. Free online English dictionary. http://dictionary.reference.com/.

Fellbaum, Christine (ed.). 1998. *WordNet: An electronic lexical database*. MIT Press.

Fellbaum, Christine. (1998). Towards a representation of idioms in WordNet. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems [online]* Montreal.

Flickinger, Dan. 2011. Accuracy vs. robustness in grammar engineering. In Bender, E. M., & Arnold, J. E. (eds.) *Language from a Cognitive Perspective: Grammar, Usage, and Processing*(pp. 31-50). CSLI Publications: Stanford.

Miyazaki, Masahiro, Satoru Ikehara & Akio Yokoo. 1993. Combined word retrieval for bilingual dictionary based on the analysis of compound word. *Transactions of the Information Processing Society of Japan* 34(4). 743–754. (in Japanese).

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Nunberg, Geoffrey, Ivan A. Sag & Tom Wasow. 1994. Idioms. *Language* 70. 491–538.

Richter, Frank & Manfred Sailer. 2009. Phraseological clauses in constructional HPSG. In Stefan Müller (ed.), *Proceedings of the 16th international conference on head-driven phrase structure grammar, university of göttingen, germany*, 297–317. Stanford: CSLI Publications. http://cslipublications.stanford.edu/HPSG/2009/.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk (ed.), *Computational linguistics and intelligent text processing: Third international conference: Cicling-2002*, 1–15. Hiedelberg/Berlin: Springer-Verlag.

Zhang, Yi, Valia Kordoni, Aline Villavicencio & Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the workshop on multiword expressions: Identifying and exploiting underlying properties*, 36–44. Sydney, Australia: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W06/W06-1206.

[1] Creative Commons Attribution Only